DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
38

DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS
38

FITTING SETS TO PROBABILITY DISTRIBUTIONS

MEELIS KÄÄRIK

TARTU UNIVERSITY
PRESS

Faculty of Mathematics and Computer Science, University of Tartu, Tartu

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (Ph. D.) in mathematical statistics on December 17, 2004, by the Council of the Faculty of Mathematics and Computer Science, University of Tartu

Supervisor:
Professor, Cand. Sc. Kalev Pärna
University of Tartu
Tartu, Estonia

Opponents:
Professor, Dr. Hans-Hermann Bock
RWTH Aachen University
Aachen, Germany

Professor, Dr. habil. Mindaugas Bloznelis
Vilnius University
Vilnius, Lithuania

The public defence will take place on January, 28, 2005.

Publication of the dissertation is financed by the Institute of Mathematical Statistics, University of Tartu (ESF grants 3963 and 5277 and research project TMTMS1776).

ISBN 9949-11-006-8 (trükis)
ISBN 9949-11-007-6 (PDF)

Tartu, Ülikooli Kirjastus
www.tyk.ut.ee
Tellimus nr. 700

Magnusele ja Maikele

Contents

Acknowledgements

I wish to express deep gratitude to my supervisor Professor Kalev Pärna for his guidance, encouragement and many helpful discussions during the completion of this work.

I am grateful to Dr. Jüri Lember for many new ideas, helpful comments and suggestions.

I am also thankful to my teachers and colleagues in the Institute of Mathematical Statistics for their support.

I am strongly indebted to all teachers in the Faculty of Mathematics and Computer Science for the help and knowledge they have given during my studies.

I am very grateful to my family for understanding, patience and moral support.

The work is partially supported by Estonian Science Foundation grants 3963 and 5277.

List of original publications

1. Käärik, M. (2000). Approximation of distributions by sphere. *Multivariate Statistics. New Trends in Probability and Statistics. Volume 5.* VSP/TEV, Vilnius-Utrecht-Tokyo, 61–66.

2. Käärik, M., Pärna, K. (2003). Approximation of distributions by parametric sets. *Acta Applicandae Mathematicae,* **78**, 175–183.

3. Käärik, M., Pärna, K. (2004). Fitting parametric sets to probability distributions. *Acta et Commentationes Universitatis Tartuensis de Mathematica. Vol. 8*, 101–112.

Introduction

The problem discussed in this work is a far-going generalization of an archetypical problem - how to approximate a distribution by a single point. For example, the expectation EX of a random variable X is the best approximation of X, since it minimizes the quadratic loss-function $E(X-a)^2$ over all possible choices of $a \in \mathbb{R}$. On the other hand, the median of a distribution on the real line is the best one-point approximation w.r.t. the loss-function $E|X-a|$.
Approximation of distributions by certain sets or models is, in fact, the key problem of several probabilistic and statistical methods. It is natural to minimize the information loss, caused by the model, in the sense of a chosen loss-function. (The term *loss* in the probabilistic context was first introduced by Wald (1939)).

The general setup of fitting sets to distributions is the following. Given a probability measure P on a separable metric space (S, d) and a class $\mathcal{A}$ of subsets of S, find an approximative set $A \in \mathcal{A}$ which minimizes the loss-function

$$W(A, P) = \int_S \varphi(d(x, A))P(dx),$$

where $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ is a nondecreasing discrepancy function and $d(x, A) = \inf\{d(x, a) : a \in A\}$.

Besides two simple examples of one-point approximations (mean and median), a variety of more complex sets are of interest in several theoretical and practical fields. Extension of the class of approximative sets to finite sets consisting of k points leads us to the concept of *k-centres* (sometimes called k-means, principal points), which is the most studied type of approximative sets during last decades (Pollard, 1981; Cuesta and Matrán, 1988, 1989; Lember, 1999). In practice, k-centres are widely used in signal

discretization or *quantization* (see, e.g., Graf and Luschgy, 2000; Gray and Neuhoff, 1998) and classification (Bock, 1996).

Further on, approximations via lines or planes is a standard idea in classical statistical methods. For example, orthogonal linear regression analysis of two random variables X and Y can be considered as search for a line which is closest (in average) to the random point $(X, Y) \in \mathbb{R}^2$. Similarly, the objective of principal component analysis is to find a k-dimensional linear manifold that comes closest to a given data set in the sense of mean squared error (see, e.g., Kendall and Stuart, 1966, Chapter 43).

In addition to classical linear methods, non-linear approximation techniques have been studied in last decades (Hastie and Stuetzle, 1989; Tibshirani, 1992). Naturally, sometimes curves and surfaces can give a better fit to the distribution, although they are more difficult to find in practice. With the increasing computational power available, the use of non-linear methods is becoming more and more common. The problem of fitting circles, spheres, ellipses, hyperbolas or other algebraic spherical shapes to a given probability distribution is an actual issue in aircraft industry, metrology, astronomy and sound waves propagation (Späth, 1996).

The aim of this work is to achieve a new level of generality of approximative sets that would cover a wide spectrum of fitting problems in practice. More precisely, we will focus on the following two classes of approximative sets: 1) finite unions of sets with uniformly bounded diameter and 2) the class of (multiple) parametric sets. The first class contains, among others, sets of k points, unions of k spheres and other, more complicated bounded geometrical shapes. Parametric representation of approximative sets is useful in many cases, where it gives economy in description: lines, planes, hyperplanes, parametric curves and surfaces. Clearly, the classes 1) and 2) intersect, since some bounded sets can also be treated parametrically.

The approximation of distributions offers many theoretical and practical (including computational) problems. The problems begin with the very existence of optimal approximative sets: even for the real line $\mathbb{R}$ the existence of the expectation EX of a random variable X is not guaranteed. In this work we search for answers to the following questions:

1) when does an optimal approximative set exist (in a given class $\mathcal{A}$),

2) does the sequence of infimum values of the loss-function converge,

3) does the sequence of optimal approximative sets converge.

The two latter questions arise when the approximative sets are found for a sequence of measures $\{P_n\}$ which converges weakly to the initial measure P, i.e. $P_n \Rightarrow P$. This is an important issue in the situation where the distribution P itself is not known (or hard to measure) but the distributions P_n are known. The situation is very common in statistics where we usually do not know the distribution of the population and only corresponding empirical distributions P_n are available. Then the convergence of empirically optimal approximating sets to an optimal set for the parent distribution is a very desirable property called *consistency*. Let us recall that the consistency of empirical means is a fundamental fact, the Law of Large Numbers. The consistency of empirical k-centres has also been a subject of various studies. Pollard (1981) proved the consistency of empirical k-centres in $\mathbb{R}^d$ for a wide class of discrepancy functions, assuming the uniqueness of the k-center. Cuesta and Matrán (1988, 1989) showed that this result also holds for uniformly convex Banach spaces. These problems were further analyzed by Lember (1999), who proved the convergence of empirical k-centres in a wide class of spaces (e.g. in reflexive Banach spaces and in Jefimov-Stečkin spaces) without the restriction of uniqueness. However, all these results are concerned only with k-point sets and empirical measures. In this work, we will extend the consistency problem to a more general level: instead of just empirical measures we explore a rather broad class of weakly converging sequences $\{P_n\}$, including those generated by stationary processes. Therefore we will achieve much wider applicability of our results.

Note that, in practice, the problem of fitting sets to distributions also contains algorithmic issues like computational effectiveness and convergence of iterative algorithms. A classical reference here is Lloyd (1988). However, these aspects are not considered in this work – we keep track on purely analytical methods.

The thesis is organized in the following way.
Chapter 1 explains the problem of approximation of distributions by sets and gives an overview of previous results in this area. Starting with the definition and basic properties of the loss-function, the chapter leads to the formulation of consistency problem. Then a more general convergence problem of optimal approximations for arbitrary weakly converging sequence of

measures is raised. Related concepts of Vapnik's statistical learning theory are also introduced and some results from this field are given.

Chapter 2 provides the reader with basic tools and assumptions we use in our analysis. First, some fundamental theorems from the theory of weak convergence are given and the importance of uniform convergence of expectations is stressed. Then assumptions about the discrepancy function φ and the measures P and $\{P_n\}$ are stated, and several types of sequences $\{P_n\}$ are shown to meet these assumptions: empirical measures, certain conditional measures, and measures generated by ergodic processes. While the empirical measures have been analyzed in this context before (Pärna, 1988), the last two examples are new.

In Chapter 3 we prove some general results concerning the loss-function, the main result being the convergence theorem of the infimum values of the loss-function, $W(P_n) \to W(P)$, as $P_n \Rightarrow P$, in separable metric spaces (Theorem 3.1.11). This theorem generalizes previous results obtained for k-centres (see, e.g., Pollard, 1981; Abaya and Wise, 1984; Pärna, 1988; Lember, 1999) to arbitrary class of approximative sets.
As an important step, Theorem 3.1.10 establishes the uniform convergence of the loss-function over a class of approximative sets that intersect with a given ball $\bar{B}(x_0, R)$ (we call such a class *uniformly visible*):

$$\lim_n \sup_{\substack{A\in\mathcal{A} \\ A\cap\bar{B}(x_0,R)\neq\emptyset}} |W(A, P_n) - W(A, P)| = 0.$$

One of our main objectives, the convergence of infimum values for the loss-function, is proved in Theorem 3.1.11.
Furthermore, properties of optimal approximative sets consisting of k component-sets are also studied in Chapter 3. The main result for such class of sets is given in Theorem 3.2.4: provided that $\epsilon > 0$ is small enough, there exists a ball in S, which for every ε-minimizing sequence for $W(\cdot, P)$ intersects with each component-set, eventually. The last result is used in next two chapters to answer the questions 1) and 3) about the existence and convergence of optimal approximative sets.

In Chapter 4 the class of multiple bounded sets with diameter not exceeding a fixed constant is introduced as a class of approximative sets, and the space of this type of sets is metrized via the Hausdorff metrics. The existence of optimal approximative sets in finite-dimensional normed space S is proved

in Theorem 4.4.3. The convergence of ε-optimal approximative sets for P_n to the class of ε-optimal approximative sets for P is stated as Theorem 4.4.4. These two results together imply the convergence of optimal approximative sets, as stated in Theorem 4.4.5. An extension where the approximative set is a union of k component-sets from different K-bounded classes $\mathcal{A}_i$ is also covered in Chapter 4 (see Theorem 4.5.2). The results of Chapter 4 generalize those of Pollard (1981), Abaya and Wise (1984), Averous and Meste (1997), Pärna, Lember and Viiart (1999).

In Chapter 5 fitting multiple parametric sets to distributions is considered. The chapter starts with the definition of 'good' parametrization and examples of proper parametrization for some classes of sets are included. Properties of optimal approximative sets are now obtained indirectly, via optimal values of parameters. We obtain several convergence results, provided that S is a separable normed space and parameter space T is a finite-dimensional normed space. To avoid technical difficulties, we first consider the case when all the component-sets belong to the same class. The main results are the existence of optimal values of parameters (Theorem 5.4.3) and the convergence of ε-optimal values of parameters for P_n to the class of ε-optimal values of parameters for P (Theorem 5.4.4). The convergence of optimal values of parameters follows immediately (Theorem 5.4.5). Same results are valid for the case when approximative component-sets are not necessarily from the same class as stated in Theorem 5.5.1.

Most of the results given in Chapters 3, 4 and 5 are published in Käärik (2000) and Käärik and Pärna (2003, 2004). In some cases the results are stated in a more general form than in our published papers. When there is no reference in the text, the result is new.

Chapter 1

The problem of fitting sets to probability distributions

In this chapter we formulate the problem of approximation of probability distributions by sets as the problem of minimization of a loss-function and indicate some arising basic theoretical questions. An overview of the literature is given to study and compare the results obtained by different authors.

1.1. The loss-function

Our objective is to approximate a given probability distribution (or data) with a set from a given class. To do this, we first have to specify how to measure the quality of approximative sets.

Let us have a separable metric space S with metrics d on it. Let $\mathcal{B}$ be the Borel σ-algebra generated by open subsets of S, and let P be a probability measure on the measurable space $(S, \mathcal{B})$. We approximate the probability distribution P by a set A from a class $\mathcal{A}$, a class of subsets of S, in the sense that 'good' A contains points from areas of high probability while the regions with small probability masses are allowed to be not represented in A. In this chapter we do not put any constraints on the class $\mathcal{A}$, however, some general assumptions about $\mathcal{A}$ will be introduced in later chapters.

Definition 1.1.1. We measure the quality of approximation of the distri-

bution P by a set A via the following *loss-function*:

$$W(A,P) := \int \varphi(\inf_{a\in A} d(x,a))P(dx), \tag{1.1}$$

where $A \in \mathcal{A}$ and φ is a nondecreasing function, called *discrepancy function* (sometimes also *penalty function*).

The full list of requirements (F1–F5) on φ will be introduced in Chapter 2. Note that a popular choice of φ is the quadratic function $\varphi(x) = x^2$.
Our loss-function can also be expressed in terms of random elements. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $X : \Omega \to S$ be a random element in S, having distribution P, $X \sim P$. Then we have

$$W(A,P) = W(A,X) \equiv E\varphi(\inf_{a\in A} d(X,a)),$$

where E means expectation.
The infimum value of the loss-function is denoted by $W(P)$, i.e. $W(P) = \inf_{A\in\mathcal{A}} W(A,P)$.

Definition 1.1.2. Any A^* satisfying $W(A^*,P) = W(P)$ is called *optimal approximative set* for P.

Definition 1.1.3. Any A^ε satisfying $W(A^\varepsilon,P) \le W(P) + \varepsilon$ is called *ε-optimal approximative set* for P.

The class of optimal approximative sets is denoted by $\mathcal{U}(P)$ and the class of ε-optimal approximative sets by $\mathcal{U}^\varepsilon(P)$.

1.2. Classes of approximative sets

We now give some examples of classes of approximative sets that have been studied in the literature.

Example 1.2.1 (φ-means)**.** Let $\mathcal{A}$ consist of 1-point subsets of S. Then optimal approximative sets are those $a^* \in S$ which minimize $E\varphi(d(x,a))$. We call them *φ-means* of P (see also Cuesta-Albertos, Gordaliza and Matrán, 1998). Note that the expectation EX itself is a special case of φ-means (choose $S = \mathbb{R}^d$ and $\varphi(d(x,a)) = \|x-a\|^2$).

Example 1.2.2 (k-centres)**.** If approximative sets consist of at most k points of S, $\mathcal{A} = \{A \subset S : |A| \leq k\}$, then optimal k-sets are called *k-centres*. This approach is useful e.g. in discretization of a continuous random variable X, which is to be transmitted through a discrete channel capable of admitting k distinct values only. The process of discretization of X is called *quantization* and many probabilistic, statistical and computational problems of quantization have been investigated by a number of authors (see, e.g., Graf and Luschgy, 2000; Gray and Neuhoff, 1998). Well known special cases of k-centres in $S = \mathbb{R}^d$ are k-means and k-medians, obtained by specifying $\varphi(\inf_{a_i \in A} d(x, a_i))$ as $\min_{a_i \in A} \|x - a_i\|^2$ or $\min_{a_i \in A} \|x - a_i\|$, respectively.

Example 1.2.3 (Orthogonal linear regression)**.** Consider the space $S = \mathbb{R}^2$. The objective of orthogonal linear regression is to find a line which is closest (in average) to the random point $(X, Y) \in \mathbb{R}^2$ (Anderson, 1984). This problem can be stated as finding an optimal line

$$l(a, b) = \{(x, y) : y = a + bx,\ (x, y) \in \mathbb{R}^2\}$$

from the class of all lines $\mathcal{A} = \{l(a, b) : a, b \in \mathbb{R}\}$.

Example 1.2.4 (Principal component analysis)**.** A related problem of principal component analysis is to find, for a given random vector X in $\mathbb{R}^d$, a k-dimensional linear manifold which is closest (w.r.t. quadratic loss) to X (Kendall and Stuart, 1966, Chapter 43). Let $M(a_{i,j})$ be a manifold consisting of all elements $(x_1, \ldots, x_d) \in \mathbb{R}^d$ which satisfy $d - k$ equations

$$\begin{aligned} a_{1,0} + a_{1,1}x_1 + \ldots + a_{1,d}x_d &= 0, \\ &\ldots \\ a_{d-k,0} + a_{d-k,1}x_1 + \ldots + a_{d-k,d}x_d &= 0. \end{aligned}$$

Then the class $\mathcal{A}$ of our interest contains all k-dimensional manifolds:

$$\mathcal{A} = \{M(a_{i,j}) : a_{i,j} \in \mathbb{R}, i = 1, \ldots, d - k; j = 0, 1, \ldots, d\}.$$

Example 1.2.5 (Principal component clustering)**.** In the cluster analysis, the objective is to detect groups of similar objects (clusters) in the data. Classical k-means and k-median clustering are well-known, but often some additional information about the given data may offer more suitable choices of approximative sets. For example, if the data is concentrated around

specific linear manifolds, the *principal component clustering* will fit the data better than the k-means model. The objective of this method is to find an optimal system of linear manifolds $\mathcal{M} = (M_1, \ldots, M_k)$ from the class $\mathcal{A}$ of all such systems. Clustering problems of this kind are studied (among many other classification problems) in Bock (1996).

Example 1.2.6 (Second order curves)**.** For $S = \mathbb{R}^2$ also circles, ellipses and hyperbolas are important classes of approximative sets. Effective procedures for calculation of optimal second order curves have been developed in Späth (1996, 1997abc), Chernov and Ososkov (1984) and Chernov and Lesort (2004). A related concept of principal curves and surfaces is explained in subsection 1.4.2.

Example 1.2.7 (Balls)**.** Approximation of distributions by a ball with a given radius was treated in Averous and Meste (1997). In Pärna, Lember and Viiart (1999) the problem of fitting several balls was studied.

Our motive is to work with general classes of approximative sets rather than to study various special cases separately. We have found it useful to analyze approximation by two wide classes of sets: bounded sets and parametric sets.

More precisely, our first interest is to approximate probability distributions by unions of bounded sets, i.e. by the elements of the class

$$\mathcal{A} = \{A^1 \cup \ldots \cup A^k : \Delta(A^i) < K, i = 1, \ldots, k\},$$

where Δ denotes the diameter and k, K are given. The advantage of using bounded sets is that the space of bounded sets can be metrized via Hausdorff metrics and thus the properties of approximative sets can be studied directly.

Secondly, as we have seen from the examples above, the approximative sets can sometimes be suitably parameterized. In that case it can be easier to study properties of optimal parameters Θ instead of approximative sets $A(\Theta)$ themselves. The general form of this class of approximative sets is given by

$$\mathcal{A} = \{A(\Theta) : \Theta \in T\}$$

with T being the space of values of the parameter Θ.

The classes of bounded and parametric sets are wide enough to include all the examples given above in this section. At the same time, they cover many other potential classes of approximative sets, undiscussed so far. For example, line segments, triangles, balls, polyhedrons, chains etc. can be useful instruments to approximate the data.

1.3. Basic problems in fitting sets to distributions

1.3.1. Existence of optimal sets

In this work we are interested in approximation of distributions in general functional spaces S. Topological properties of S are closely related to the problems of existence and convergence of optimal approximative sets. Measure-theoretical considerations motivated us to focus on separable metric spaces. As will be seen in Chapter 3, several important properties of the loss-function can be proved in separable metric spaces. Unfortunately, separability is not a sufficient condition for the existence of an optimal approximative set. A simple counterexample can be constructed to demonstrate that.

Example 1.3.1. Consider a 1-center approximation of a zero-mean distribution P on $\mathbb{R}$ (a separable metric space) and discrepancy function $\varphi(x) = x^2$. Clearly, the optimal 1-center (the expectation EX, $X \sim P$) lies at 0. Consider now the space $\mathbb{R}\backslash\{0\}$, which is separable as well, but it does not contain the expectation of the conditional measure $P(\cdot|\mathbb{R}\backslash\{0\})$.

In this example, of course, the space $\mathbb{R}\backslash\{0\}$ is not complete. But the next example shows that even in a complete space optimal approximative sets need not exist.

Example 1.3.2. Consider the space c_0 – the space of all sequences converging to zero. It is a Banach space, equipped with norm $\|x\| = \sup_{1\leq k<\infty} |\xi_k|$, $x = (\xi_k) \in c_0$. Let us have a random variable X with distribution P concentrated at the elements of c_0 by following:

$$
\begin{aligned}
x_1 &= (2,0,0,\ldots), \quad P(x_1) = \tfrac{1}{2};\\
x_2 &= (0,2,0,\ldots), \quad P(x_2) = \tfrac{1}{4};\\
&\ldots\\
x_k &= (\overbrace{0,\ldots,0,2}^{k},0,\ldots), \quad P(x_k) = \tfrac{1}{2^k};\\
&\ldots
\end{aligned}
$$

The objective is to find the optimal 1-center $a = (a_1, a_2, \ldots)$ corresponding to the quadratic discrepancy function $\varphi(x) = x^2$. Let us evaluate the loss-function

$$
\begin{aligned}
W(a,P) &= \sum_{i=1}^{\infty} \|a - x_i\|^2 P(x_i) = \sum_{i=1}^{\infty} \sup_{1\le k<\infty} (a_k - (x_i)_k)^2 P(x_i)\\
&\ge \sum_{i=1}^{\infty} (a_1 - (x_i)_1)^2 P(x_i) = \frac{1}{2}(a_1-2)^2 + \frac{1}{2}(a_1-0)^2\\
&= a_1^2 - 2a_1 + 2. \qquad (1.2)
\end{aligned}
$$

The polynomial at the right hand side attains its minimum at $a_1 = 1$, so the infimum value of loss-function cannot be less than 1. As seen from the formula (1.2), the value 1 can be obtained only if for each $i = 1, 2, \ldots$ the equality $\sup_{1\le k<\infty}(a_k - (x_i)_k)^2 = (1 - (x_i)_1)^2 = 1$ holds. But this requirement is only satisfied by $a = (1, 1, \ldots)$, which is not the element of c_0.

The problem of existence of k-centres in terms of topological properties of S is thoroughly studied in Lember (1999). We bring two results from there.

Theorem 1.3.3. *Let P be a probability measure on space S and let τ be a topology on S such that every closed ball in S is sequentially τ-compact*[1]*. Then the class of k-centres for P is not empty.*

Corollary 1.3.4. *Let P be a probability measure on space S, which is one of the following spaces:*

1. *finite dimensional space;*

[1] A set V is called sequentially compact, if every sequence in V has a converging subsequence (to $v \in V$).

2. reflexive Banach space;

3. a dual of a separable linear space.

Then the class of k-centres for P is not empty.

In the current work we study the existence problem for much more general approximative sets as compared to k-centres. In Chapters 4 and 5, the existence of optimal bounded and parametric sets will be proved for finite-dimensional normed spaces which cover majority of practical applications.

1.3.2. Convergence of optimal sets and infimum values of the loss-function

In practice, it often happens that the measure P itself is not known and only a measure P_n (which is close to P) is available. Therefore, we are forced to optimize $W(A, P_n)$ instead of $W(A, P)$, and a natural question is, whether the P_n-optimal set A_n^* is close to the P-optimal set A^*.
This is the typical situation in statistics where P is replaced by its empirical measure P_n, defined as follows. Let $x_1, x_2, \ldots, x_n$ be independently sampled from P (in space S), and let P_n be defined by $P_n(B) = \frac{\#\{x_i \in B\}}{n}$ for every measurable $B \subset S$. The measure P_n is called *empirical measure* corresponding to P.
If, at the same time, an estimator based on P_n (in our case A_n^*) converges in probability to corresponding theoretical value (in our case A^*), then the estimator is called *consistent*. We speak about *strong consistency*, if the sequence of estimators converges with probability 1.

So far, the consistency of empirical approximative sets has been studied basically for k-centres, only (see, e.g., Pollard (1981) and other papers cited in subsection 1.4.1). In this work we cover 1) much wider classes of approximative sets (see Section 1.2), and 2) wider class of measures $\{P_n\}$.

As to 2), according to Varadarajan (1958), the sequence of empirical measures converges with probability 1 weakly to the parent measure P, i.e.

$$\mathbf{P}\{P_n \Rightarrow P\} = 1.$$

Therefore, it is natural to consider a more general convergence problem where the measures P_n are not restricted to be empirical. Let $P_1, P_2, \ldots$ be

arbitrary probability measures on the space S satisfying $P_n \Rightarrow P$ (we will give several examples of such measures in Chapter 2).

We search answers for the following questions:

a) Provided that $P_n \Rightarrow P$, does the sequence of infimum values of the loss-function for $\{P_n\}$ converge to that of loss-function for P when n tends to infinity, i.e. does the convergence

$$W(P_n) \to W(P),\ n \to \infty,$$

take place?

b) Provided that $P_n \Rightarrow P$, do the ε-optimal approximative sets A_n^ε for P_n converge (in Hausdorff metrics h) to the class of ε-optimal approximative sets for P, i.e. does the convergence $h(A_n^\varepsilon, \mathcal{U}^\varepsilon(P)) \to 0$ take place, if $n \to \infty$?

In cases when an optimal approximative set exists, the answer to *b*) for $\varepsilon = 0$ also gives the answer to the following question of special interest: do optimal approximative sets A_n for P_n converge to the class of optimal approximative sets for P, i.e. does $h(A_n, \mathcal{U}(P)) \to 0$, if $n \to \infty$? In Chapters 3, 4 and 5 positive answers to these questions will be given.

1.4. Overview of the literature

1.4.1. The problem of k-centres

Sets consisting of k points are the most studied case of approximative sets. In applications, it is common that a continuous random variable is to be approximated by a discrete random variable with finite number (k) of possible values. For example, transmission of continuous signal through a channel with finite number of possible states is an important task of this kind.
When approximating by k points, the loss-function can be decomposed as

$$W(A,P) = \int \min_{a_i \in A} \varphi(d(x,a_i))P(dx) = \sum_{i=1}^{k} \int_{C_i} \varphi(d(x,a_i))P(dx),$$

where $\{C_1, \ldots, C_k\}$ is a Voronoi partition of S, generated by k points $a_1, \ldots, a_k$. For each $i = 1, \ldots, k$, the Voronoi region C_i acquires all elements of S whose closest element in A is a_i (the borders are classified to regions with minimum index).

A pioneering work in the field is MacQueen (1967). MacQueen investigated a random variable X with distribution P in N-dimensional Euclidean space and a sample $x_1, x_2, \ldots$ from P. He provided an online algorithm, where initial centres are chosen randomly from the data points, and then adjusted iteratively by assigning the data samples to the nearest clusters and recomputing the centres. Quadratic loss-function was chosen to measure the optimality of a center. The almost sure convergence of certain variables, related to 'running' k-centres, and the convergence of infimum values of the loss-function were proved. Mainly the tools from the theory of martingales were used to achieve these results. The algorithm is quite simple and easy to use, but will not necessarily lead to the global optimum, and it is also sensitive to the initial choice of cluster centres.

In Hartigan (1978) partitions generated by k points on the real line were investigated. The almost sure convergence of empirically optimal cutting points to theoretical values and the asymptotic normality of empirically optimal cutting points were proved.

H. Sverdrup-Thygeson investigated 1-centres defined by power type discrepancy function, in compact metric spaces (Sverdrup-Thygeson, 1981). Optimal points were called centroids, and he also studied properties of 'restricted' centroids – best points among the sample elements. The convergence of empirical centroids and empirical restricted centroids was proved.

D. Pollard studied k-centres in $S = \mathbb{R}^d$ for more general discrepancy functions (Pollard, 1981). He assumed that φ satisfies $\varphi(0) = 0$, $\varphi(r) \to \infty$ $(r \to \infty)$, and it has Δ_2-property, i.e. $\exists \lambda : \quad \varphi(2r) \leq \lambda \varphi(r), \quad \forall r > 0$. He proved the convergence of empirical k-centres in Hausdorff metrics, assuming that the optimal k-center is unique.

E. E. Abaya and G. L. Wise also focused on the convergence problem of k-centres in $\mathbb{R}^d$. They demonstrated that the convergence of infimums of the loss-function takes place for more general measures as compared to empirical measures (Abaya and Wise, 1984). Namely, it holds for any sequence $\{P_n\}$ converging weakly to P, provided that the discrepancy function φ is uniformly integrable w.r.t. $\{P_n\}$.

J. A. Cuesta and C. Matrán, using the Skorokhod representation theorem (instead of commonly used uniform SLLN), focused on the sequence of P_n-distributed random variables X_n rather than the sequence $\{P_n\}$ it-

self (Cuesta and Matrán, 1986). They showed convergence of empirical k-centres in uniformly convex Banach spaces (Cuesta and Matrán, 1988, 1989). Uniformly convex Banach spaces have some useful properties (like reflexivity) to solve problems of this kind. Although a wide class of discrepancy functions was considered, the restriction of uniqueness of k-centres still remained. This question was addressed in Cuesta and Matrán (1988), where the uniqueness is shown to be an important condition for the convergence of k-centres.

K. Pärna considered k-centres in separable metric spaces without assuming uniqueness of the optimal set (Pärna, 1986). The almost sure convergence of empirical infimums of loss-function to the theoretical one was proved. He also studied the properties of the loss-function for general measures $\{P_n\}$ introduced in Abaya and Wise (1984), and proved that the convergence of infimums of the loss-function for such measures takes place for k-centres in separable metric spaces as well (Pärna, 1988).

The existence and convergence problems of k-centres in reflexive Banach spaces are studied in Pärna (1990). The assumption of uniqueness of k-centres is dropped and the main result states that the distance between empirically optimal k-centres and the set of theoretically optimal k-centres tends to zero almost surely.

J. Lember gives a systematic account of k-centres theory (Lember, 1999). The convergence of empirically optimal k-centres is proved in various normed spaces, including reflexive Banach spaces and Jefimov-Stečkin spaces. Lember also revealed a connection between k-centres and best approximation theory, a branch of functional analysis. Consistency properties of k-centres are further analyzed in Lember and Pärna (1999) and Lember (2002) and arbitrary minimizing sequences are thoroughly explored in Lember (2003).

Yet another interesting approach to k-centres is given by B. Flury (1990, 1993), who investigated the properties of k-centres (called *principal points* by him) of univariate symmetric distributions, univariate and bivariate normal distribution and multivariate distributions. He revealed a relationship between principal points and principal components for the case of $k = 2$ for normal and elliptical distributions.

1.4.2. Fitting more general sets

Besides k-centres more general approximations have been investigated. J. Averous and M. Meste treated the 'median balls' problem, where approximation by a ball with a given radius in $S = \mathbb{R}^d$ was studied, using discrepancy function $\varphi(x) = |x|$ (Averous and Meste, 1997). The existence of median balls and the convergence of median balls of P_n to median balls of P was proved, provided that $P_n \Rightarrow P$ and S satisfies certain convexity conditions.

The approximation of distributions by optimal balls in finite-dimensional normed spaces was studied in Pärna, Lember and Viiart (1999). Using Skorokhod representation theorem, the convergence of empirically optimal balls to the set of optimal balls for the theoretical distribution P was proved.

The author of this work studied approximation of distributions by spheres with fixed radius in finite-dimensional Euclidean space (Käärik, 2000). Some consistency results were obtained by using techniques from Pollard (1981).

Next we explain a related concept of principal curves.

As an extension to principal components, T. Hastie and W. Stuetzle (1989) introduced *principal curves* as smooth curves that are *self-consistent* for a distribution or data set. The basic idea is to find a smooth one-dimensional curve that passes through the 'middle' of a p-dimensional data set, providing a nonlinear summary of the data. An algorithm for constructing such curves is provided. The method was successfully used to improve locations of 950 magnets of Stanford linear collider with the aim to achieve better focused beam of particles.

The property of *self-consistency*, first introduced in connection with principal curves, is further studied in Tarpey and Flury (1996). A random vector Y is called self-consistent for X if each point of the support of Y is the conditional mean of X, given that X projects onto that point. This approach allows to unify the theoretical basis for several statistical methods, e.g. principal curves and surfaces, principal points, principal modes of variation.

B. Kégl, et al. (2000) took another approach to principal curves by defining them as continuous curves of a given length which minimize the expected

squared distance between the curve and a random point. Note that this approach can ideally be embedded into our approximation problem. The authors provided a learning scheme for finding fixed length polygonal lines with k-segments, and analyzed convergence properties of the method.

Another alternative interpretation of principal curves is given by R. Tibshirani (1992), defining them as curves minimizing a penalized log-likelihood measure. The ideas in this direction are further developed by J. J. Verbeek, N. Vlassis and B. Kröse (2000, 2001).

1.5. Relationship with Vapnik's problem of empirical risk minimization

In this subsection we compare our approach with the empirical risk minimization problem in statistical learning theory (see, e.g., Vapnik, 1998, Part I). We give a short overview of Vapnik's theory in order to see similarities and differences between the two approaches.

The loss-function considered in statistical learning theory has the form

$$W(\alpha) = \int \phi(z, \alpha) dP(z) \to \min_{\alpha \in \Lambda},$$

where P is a probability distribution on $\mathbb{R}^d$ and α is a parameter.

Given the i.i.d. data $z_1, \ldots, z_n$ from P we define the empirical loss-function by

$$W_{emp}(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \phi(z_i, \alpha),$$

and the optimal value of α will be denoted by $\alpha_n = \alpha(z_1, \ldots, z_n)$, i.e. $W_{emp}(\alpha_n) = \min_{\alpha \in \Lambda} W_{emp}(\alpha)$.

Definition 1.5.1. The principle of empirical loss minimization is said to be consistent for the functions $\{\phi(z, \alpha),\ \alpha \in \Lambda\}$ and for the probability distribution function $P(z)$, if the following sequences converge in probability to same limit:

$$W(\alpha_n) \to \inf_{\alpha \in \Lambda} W(\alpha)$$

and

$$W_{emp}(\alpha_n) \to \inf_{\alpha \in \Lambda} W(\alpha).$$

It turns out that this classical definition is too weak to exclude some trivial cases of consistency. Therefore Vapnik defines strict consistency.

Definition 1.5.2. The method of minimizing empirical loss is strictly (nontrivially) consistent for the set of functions $\{\phi(z,\alpha),\ \alpha \in \Lambda\}$ and for the probability distribution function $P(z)$, if for any nonempty subset $\{\Lambda(c),\ c \in (-\infty,\infty)\} \subset \Lambda$ where

$$\Lambda(c) = \left\{\alpha :\ \int \phi(z,\alpha)dP(z) \geq c\right\}$$

the convergence

$$\inf_{\alpha\in\Lambda(c)} W_{emp}(\alpha_n) \xrightarrow{\mathbf{P}} \inf_{\alpha\in\Lambda(c)} W(\alpha),\ n \to \infty$$

is valid (in probability).

We will next present the 'key' theorem of empirical loss minimization, where an equivalent (and more convenient) condition for the consistency of an empirical loss minimization principle has been proposed (Vapnik, 1998, Theorem 3.1).

Theorem 1.5.3 (The key theorem of statistical learning)**.** *Let there exist constants a and A such that for all functions in the set $\{\phi(z,\alpha),\ \alpha \in \Lambda\}$, and for a given distribution function $P(z)$, the inequalities*

$$a \leq \int \phi(z,\alpha)dP(z) \leq A$$

hold true. Then the following statements are equivalent:

1. *For the given distribution function $P(z)$ the empirical loss minimization principle is strictly consistent on the set of functions $\{\phi(z,\alpha),\ \alpha \in \Lambda\}$.*

2. *For the given distribution function $P(z)$ the uniform one-sided convergence*

$$\mathbf{P}\left\{\sup_{\alpha\in\Lambda}\left(\int \phi(z,\alpha)dP(z) - \frac{1}{n}\sum_{i=1}^{n}\phi(z_i,\alpha\right) > \varepsilon\right\} \to 0,\ n \to \infty$$

takes place over the set of functions $\{\phi(z,\alpha),\ \alpha \in \Lambda\}$.

The theorem states that, in order to prove that an empirical loss minimization principle is strictly consistent, it is enough to show the uniform convergence of values of the loss-function over the set of possible parameters. Furthermore, it also states that the requirement of uniform convergence is not too strong: it is not possible to construct a consistent principle if the uniform convergence fails. Sufficient and necessary conditions for uniform convergence (based on the combinatorial properties of the class of functions) are given e.g. in Vapnik (1998, Part I, Chapter 3) and Pollard (1984, Chapter 2).

Conclusion

The empirical risk minimization is related to our approximation problem, although there are several differences. The biggest difference is that the convergence of optimal α_n (for P_n) to the optimal α (for P) is not of interest in the statistical learning theory: only the quality of α_n measured by the value of Φ is important. This is also a reason why the function ϕ in Vapnik's theory has more general form than the discrepancy function we are using. In our treatment, the loss-function is specified through the distance $d(x, A)$ and this explicit form is the basis for proving the convergence of optimal approximative sets themselves. Let us remind also that we consider more general sequences of measures $\{P_n\}$ than empirical measures and cover more general spaces (separable metric space S instead of finite-dimensional Euclidean space $\mathbb{R}^d$).
As a common feature, the uniform convergence of the loss-function plays central role in both approaches.

Chapter 2

Basic tools and assumptions

In this chapter we present some mathematical tools that are needed to solve posed problems. First we bring some necessary facts from the theory of weak convergence of probability measures. Next we state general assumptions about our discrepancy function φ and measures P and $\{P_n\}$. These assumptions remain valid throughout the rest of the work. In the end of the chapter we bring some important examples of sequences $\{P_n\}$ which satisfy these conditions. Our main contribution in this chapter is Theorem 2.4.9.

2.1. Weak convergence of probability measures

Let us have a metric space S with metrics d on it. Let $\mathcal{B}$ be the Borel σ-algebra generated by open subsets of S. Let $P, P_1, P_2, \ldots$ be probability measures on $\mathcal{B}$.

Definition 2.1.1. If for every bounded and continuous function $f : S \to \mathbb{R}$ the measures $P_1, P_2, \ldots$ and P satisfy $\int_S f dP_n \to \int_S f dP$, then we say that the sequence $\{P_n\}$ *converges weakly* to the measure P.

The weak convergence will be denoted via $P_n \Rightarrow P$. Sometimes it is more convenient to use the language of random elements. Let $X, X_1, X_2, \ldots$ be random elements on S having distributions $P, P_1, P_2, \ldots$, respectively. If $P_n \Rightarrow P$ then the sequence $\{X_n\}$ is said to converge to X *in distribution* and we write $X_n \xrightarrow{D} X$.

Next we give a some basic results of the theory of weak convergence, based on Billingsley (1968).

Let h be a measurable map from the space S to the space S' (endowed with Borel σ-algebra $\mathcal{B}'$). Then the function h defines a measure Ph^{-1} on measure space $(S', \mathcal{B}')$ by the equality $Ph^{-1}(A) = P(h^{-1}A)$, $A \in S'$.

Most interesting special case in this theory is when the space S' is real line, i.e. the function h is a measurable real function, $h : S \to \mathbb{R}$. Let D_h denote the set of all points of discontinuity of the function h. Then the following theorem holds (Billingsley, 1968, Theorem 5.2).

Theorem 2.1.2. *1. If $P_n \Rightarrow P$, then $P_n h^{-1} \Rightarrow Ph^{-1}$ for every measurable function h satisfying $P(D_h) = 0$.*

2. If for every continuous bounded real function h the convergence $P_n h^{-1} \Rightarrow Ph^{-1}$ holds, then also $P_n \Rightarrow P$.

3. If $P_n \Rightarrow P$ and h is a bounded measurable real function satisfying $P(D_h) = 0$, then $\int h dP_n \to \int h dP$.

We next examine the uniform integrability and some related concepts.

Let $Y_1, Y_2, \ldots$ and Y be random variables, i.e. real measurable functions on $(\Omega, \mathcal{F}, \mathbf{P})$.

Definition 2.1.3. Random variables Y_n are called *uniformly integrable*, if the equality

$$\lim_{r\to\infty} \sup_n \int_{|Y_n|\geq r} |Y_n| d\mathbf{P} = 0$$

holds.

In the case when each $Y_n = h(X_n)$, $X_n \sim P_n$ for some real measurable function h, the previous definition reduces to

Definition 2.1.4. A function $h : S \to \mathbb{R}$ is called uniformly integrable with respect to the sequence of measures $\{P_n\}$, if the equality

$$\lim_{r\to\infty} \sup_n \int_{|h(x)|\geq r} |h(x)| P_n(dx) = 0$$

holds.

Uniform integrability allows to deduce the convergence of expectations from the convergence in distribution.

Theorem 2.1.5. *If* $Y_n \xrightarrow{D} Y$ *and* Y_n *are uniformly integrable, then* $E(Y_n) \to E(Y)$.

The same assertion in terms of the probability distributions is:
if $P_n h^{-1} \Rightarrow Ph^{-1}$ and h is uniformly integrable w.r.t. $\{P_n\}$, then $\int h dP_n \to \int h dP$.

We bring one more useful result about uniform integrability (Durrett, 1991, p. 224).

Theorem 2.1.6. *If* $Y_n \to Y$ *a.s., then the following are equivalent:*

(i) $\{Y_n\}$ *is uniformly integrable.*

(ii) $Y_n \to Y$ *in* L^1.

(iii) $E|Y_n| \to E|Y| < \infty$.

Theorem 2.1.7 (Billingsley, 1968, Theorem 5.3)**.** *If* $Y_n \xrightarrow{D} Y$, *then the inequality* $E\{|Y|\} \leq \liminf_n E\{|Y_n|\}$ *holds.*

This assertion is equivalent to the following: if $P_n \Rightarrow P$, then $\int |y| P(dy) \leq \liminf_n \int |y| P_n(dy)$.

The following theorem (Billingsley, 1968, Theorem 5.5) and corollary are also essential for us.

Theorem 2.1.8. *For a sequence* $\{h_n\}$ *and* h *let* $D = \{x | x_n \to x \Rightarrow h_n(x_n) \to h(x)\}$. *If the convergence* $P_n \Rightarrow P$ *takes place and* $P(D) = 1$, *then also* $P_n h_n^{-1} \Rightarrow Ph^{-1}$.

The latter theorem implies a useful corollary (Lember, 1999, Corollary 3.2.2).

Corollary 2.1.9. *Let* $\{h_n, h\}$ *be a sequence of measurable functions having* $P(D) = 1$. *Suppose the sequence is bounded by a continuous function* g, *satisfying* $g(x) > 0$, $\forall x \in S$. *If the convergences* $P_n \Rightarrow P$ *and* $\int g dP_n \to \int g dP (< \infty)$ *take place, then also the convergence* $\int h_n dP_n \to \int h dP$ *takes place.*

2.2. Uniform convergence of expectations

If the sequence $\{P_n\}$ consists of empirical measures corresponding to P, then the convergence $W(A, P_n) \to W(A, P)$ takes place almost surely for each fixed A by the Strong Law of Large Numbers (SLLN). For our purposes, we must ensure this convergence to hold for a wider class of weakly converging measures $\{P_n\}$.

The situation becomes more complicated, if the set A is not fixed, but depends on n. In this case we need to use stronger tools of uniform convergence. Let A^* and A_n denote P-optimal and P_n-optimal approximative sets, respectively, and let us suppose that the uniform convergence

$$\lim_n \sup_{A \in \mathcal{A}} |W(A, P_n) - W(A, P)| = 0 \tag{2.1}$$

holds. Then simple relationships

$$W(A^*, P) - W(A_n, P_n) \leq W(A_n, P) - W(A_n, P_n) \to 0$$

and

$$W(A^*, P) - W(A_n, P_n) \geq W(A^*, P) - W(A^*, P_n) \to 0.$$

would imply the convergence of infimum values:

$$\lim_{n \to \infty} W(A_n, P_n) = W(A^*, P).$$

In the case of empirical measures P_n the uniform convergence (2.1) can be deduced from the theory of empirical processes (Gaenssler and Stute, 1979). These criteria have been used in studies of consistency properties of approximative sets (see, e.g., Pollard, 1981; Pärna, 1988, 1990).

For more general sequences $\{P_n\}$ (as compared to empirical measures) important results are given by J. Ranga Rao. The next theorem (Ranga Rao, 1962) gives sufficient conditions for the uniform convergence (over a class of functions f) of integrals $\int f dP_n \to \int f dP$. This theorem plays a central role in our proof of convergence of infimum values of the loss-function in Chapter 3.

Theorem 2.2.1. *Let Φ be a family of continuous functions on a separable metric space S satisfying the following conditions:*

i) there exists a continuous function g on space S such that for any $f \in \Phi$ and $x \in S$ the inequality $|f(x)| \leq g(x)$ holds;

ii) the family Φ is equicontinuous.

Suppose that $P_1, P_2, \ldots$ and P are probability measures on space S such that

a) the weak convergence $P_n \Rightarrow P$ takes place,

b) the convergence $\int g dP_n \to \int g dP (< \infty)$ takes place.

Then the uniform convergence

$$\lim_{n\to\infty} \sup_{f\in\Phi} \left| \int f dP_n - \int f dP \right| = 0$$

holds.

2.3. Main assumptions concerning φ, P and $\{P_n\}$

Recall that our loss-function is the mean value of the discrepancy function φ, i.e. $W(A, P) = E\varphi(d(X, A))$. Let us introduce some requirements for the discrepancy function φ and measures P and P_n.

A1. The discrepancy function φ has following properties:

F1. φ is continuous,

F2. φ is nondecreasing,

F3. $\varphi(0) = 0$,

F4. function φ has Δ_2-property:

$$\exists \lambda : \quad \varphi(2x) \le \lambda \varphi(x) \quad \forall x > 0,$$

F5. the inequality $\lim_{s\to\infty} \varphi(s) =: \varphi(\infty) > 0$ holds.

A2. for some $x_0 \in S$ we have $\int \varphi(d(x, x_0)) P(dx) < \infty$;

A3. the weak convergence $P_n \Rightarrow P$ holds;

A4. for some $x_0 \in S$ the function $\varphi(d(x, x_0))$ is uniformly integrable with respect to $\{P_n\}$.

Remark 2.3.1. The requirements *F1 – F4* are weak enough to be satisfied, for example, by any power function. The requirement *F5* is also very natural, since the approximation problem has no meaning with $\varphi(x) \equiv 0$. Assumption A2 guarantees the finiteness of the loss-function (see Lemma 3.1.2 a) for details). The requirements A3 and A4 for the sequence $\{P_n\}$ will be expounded in the next section.

2.4. Some eligible classes of $\{P_n\}$[1]

We next give some examples of the sequences $\{P_n\}$ satisfying conditions A3 and A4.

Example 2.4.1 (Empirical measures)**.** The weak convergence of empirical measures $\{P_n\}$ is shown to take place with probability 1 in Varadarajan (1958), and the property A4 for these measures is verified in Pärna (1988).

The second example demonstrates easily that the class of measures satisfying A3 and A4 is wider than the class of empirical measures.

Example 2.4.2 (Conditional measures)**.** Let us have a probability measure P on S and let φ satisfy A1, A2. Write B_n for an open ball with center $x_0 \in S$ and radius n, $B_n = B(x_0, n)$. Define the (conditional) measure P_n as follows:

$$P_n(A) = \begin{cases} \frac{P(A \cap B_n)}{P(B_n)}, & \text{if } P(B_n) > 0 \\ 0, & \text{if } P(B_n) = 0. \end{cases}$$

It is easy to see that in the process $n \to \infty$ the convergence $P(B_n) \to 1$ takes place, which implies $P_n \Rightarrow P$ and, therefore, A3. To get A4, note that if $P(B_n) > 0$, then

$$P_n(A) = \frac{P(A \cap B_n)}{P(B_n)} \le \frac{P(A)}{P(B_n)}.$$

Write n_0 for the smallest index such that the ball B_{n_0} has positive P-measure. From the simple inequality

$$\sup_n \int_{\varphi(d(x,x_0))\ge a} \varphi(d(x,x_0))P_n(dx) \le \frac{1}{P(B_{n_0})} \int_{\varphi(d(x,x_0))\ge a} \varphi(d(x,x_0))P(dx)$$

[1]This section is not a prerequisite for reading Chapters 3, 4 and 5

and using assumption A2, we get

$$\lim_{a\to\infty} \sup_n \int_{\varphi(d(x,x_0))\geq a} \varphi(d(x,x_0))P_n(dx) = 0,$$

i.e. uniform integrability A4.

Next subsection is dedicated to certain random processes that generate sequences of probability measures with properties A3 and A4. Sequences with such properties occur in various practical fields, e.g. modern information theory and economics.

2.4.1. Ergodic processes as generators of $\{P_n\}$

In this subsection we give one more motivation for assumptions A3 and A4 by introducing an important class of sequences $\{P_n\}$ with these properties. Namely, such sequences can be defined by ergodic random processes $\{Y_n\}$. Before we state our result (Theorem 2.4.9), we present some useful results from Gray (1991) and Durrett (1991), revealing connections between the convergence of sample (time) averages and the uniform integrability.

Let $Y_1, Y_2, \ldots$, be real valued random variables, defined on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

Definition 2.4.3. The random sequence $\{Y_n\}$ is called (strictly) stationary, if for any $k \geq 1$ distributions of $(Y_1, Y_2, \ldots)$ and $(Y_k, Y_{k+1}, \ldots)$ coincide.

Stationary sequences can also be considered as the transformation of a single random variable Y. Suppose that we have a measurable map $f : \Omega \to \Omega$. Defining $f^2(\omega) = f(f(\omega))$ and further $f^n(\omega) = f(f^{n-1}(\omega))$ we get measurable maps f^n for each $n = 1, 2, \ldots$ Now a random variable Y together with this system defines a random process $Y_n(\omega) = Y(f^n(\omega))$. In ergodic theory, the probability space together with the map $f : \Omega \to \Omega$ is called a *dynamic system* and denoted via $(\Omega, \mathcal{F}, \mathbf{P}, f)$.

Definition 2.4.4. The dynamic system $(\Omega, \mathcal{F}, \mathbf{P}, f)$ is called *stationary* or *invariant* if the equality

$$\mathbf{P}(f^{-1}A) = \mathbf{P}(A) \tag{2.2}$$

holds for each event $A \in \mathcal{F}$. In case when (2.2) holds we also say that the measure $\mathbf{P}$ is stationary w.r.t. f and f is *measure-preserving* map.

The next lemma shows that each stationary process can be, in fact, presented via a measure-preserving map (Shiryayev, 1980, p. 391).

Lemma 2.4.5. *For each stationary stochastic process* $\{Y_i\}$ *there exist a probability space* $(\overline{\Omega}, \overline{\mathcal{F}}, \overline{\mathbf{P}})$*, a random variable* Y *and a measure-preserving map* f *such that the distributions of* $\{Y_i\}$ *and* $\{Y(f^i)\}$ *coincide.*

To simplify notation we now focus on stationary random processes $\{Y(f^i)\}$, $i \geq 1$, defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

Let P be the distribution of Y and let measures P_n be defined by

$$P_n(B) \equiv P_n^{\omega}(B) := \frac{\#_{i=1}^n \{Y_i(\omega) \in B\}}{n} \tag{2.3}$$

for each Borel set $B \in \mathcal{B}$. Below we deduce some conditions under which P and $\{P_n\}$ satisfy A3 and A4.

Similarly to the well-known case of i.i.d. random variables, we can define sample averages (or time averages) of the process $\{Y_n\}_{n \in \mathbf{N}}$ as

$$< Y >_n (\omega) = \frac{1}{n} \sum_{i=1}^{n} Y(f^i(\omega)).$$

The convergence properties of this sequence are very important in ergodic theory. Therefore we study this sequence more closely, starting with the uniform integrability property (Gray, 1991, p. 130).

Lemma 2.4.6. *Let* Y *be* $\mathbf{P}$*-integrable and* $\mathbf{P}$ *be stationary w.r.t.* f*, then the sequences* $\{Y(f^i)\}_{i=1,2,\ldots}$ *and* $\{< Y >_i\}_{i=1,2,\ldots}$ *are uniformly integrable:*

$$\lim_{r \to \infty} \sup_i \int_{|Y(f^i)| \geq r} |Y(f^i)| d\mathbf{P} = 0$$

and

$$\lim_{r \to \infty} \sup_i \int_{|< Y >_i| \geq r} | < Y >_i | d\mathbf{P} = 0.$$

Definition 2.4.7. A dynamical system (and associated random process) is said to be *ergodic* if every invariant event has probability 1 or 0 (an event $A \in \mathcal{F}$ is called invariant, if $f^{-1}A = A$).

The following theorem is one of the key results in the theory of ergodic processes.

Theorem 2.4.8. *If a dynamical system $(\Omega, \mathcal{F}, \mathbf{P}, f)$ is ergodic and Y is $\mathbf{P}$-integrable, then the following limit holds $\mathbf{P}$-a.s.:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y(f^i) = E(Y).$$

The next theorem is our main result in this chapter.

Theorem 2.4.9. *Assume that A1 and A2 hold. Let $\{Y(f^n)\}$ be an ergodic process with $Y \sim P$ and let $\{P_n\}$ be generated by $\{Y(f^n)\}$ as in (2.3). Then the assumptions A3 and A4 hold a.s., i.e.*

$$P_n \Rightarrow P \quad a.s.$$

and

$$\lim_{r\to\infty} \sup_n \int_{\varphi(|x|)>r} \varphi(|x|) dP_n = 0 \quad a.s.$$

Proof. For A3 first note that, by Theorem 2.4.8, the convergence

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(Y(f^i)) \to EI_A(Y) = P(A) \quad a.s.$$

takes place for each $A \in \mathcal{B}$ since the indicator function I_A is measurable. Let $F_n(\omega, \cdot)$ be the probability distribution function corresponding to $P_n = P_n^\omega$. We have established that $\mathbf{P}\{\omega : \ F_n(\omega, x) \to F(x)\} = 1$, $\forall x \in \mathbb{R}$, and it remains to show that $\mathbf{P}\{\omega : \ F_n(\omega, x) \to F(x), \ \forall x \in \mathbb{R}\} = 1$ holds as well. Let $\mathbb{Q}$ denote the set of rational numbers and let D be the set of discontinuity points of F. Since D is at most countable, then $\mathbb{Q} \cup D$ is countable, therefore $\mathbf{P}(\Omega^*) = 1$, where $\Omega^* = \{\omega | \cap_{x\in\mathbb{Q}\cup D} \{F_n(\omega, x) \to F(x)\}\}$. Take arbitrary $\omega \in \Omega^*$. Then for each $x \in D$, the convergence $F_n(\omega, x) \to F(x)$ takes place by the construction of Ω^*, and for each $x \in \mathbb{R}\backslash D$ one can take upper and lower bounds to x from $\mathbb{Q}$ and use the continuity of F at x to establish the convergence. Therefore $P_n \Rightarrow P$ a.s., i.e. a sequence of measures generated by an ergodic process satisfies A3 with probability 1.

For A4 notice that, since $\{Y_n\}$ is ergodic, the continuity of φ implies that the sequence $\{\varphi(|Y_n|)\}$ is also ergodic. Assumption A3 ensures that $\mathbf{P}(\Omega^*) = 1$, where $\Omega^* = \{\omega : \ P_n \equiv P_n^\omega \Rightarrow P\}$, hence, by Skorokhod's Representation

Theorem (Billingsley, 1995, Theorem 25.6), for each $\omega \in \Omega^*$ there exist random variables $X_n^\omega \sim P_n^\omega$ and $X^\omega \sim P$ on $([0,1], \mathcal{B}, Leb)$ such that $X_n^\omega \to X^\omega$ Leb-a.s. Since φ is continuous, we also have $\varphi(|X_n^\omega|) \to \varphi(|X^\omega|)$ Leb-a.s. Further, assumption A2 ensures that $E_{Leb}\varphi(|X^\omega|) = E_{\mathbf{P}}\varphi(|Y|) = \int \varphi(|x|)dP < \infty$, hence Theorem 2.4.8 applies, i.e.

$$E\varphi(|X_n^\omega|) = \frac{1}{n}\sum_{i=1}^{n} \varphi(|Y(f^i(\omega))|) \to E\varphi(|Y|) = E\varphi(|X^\omega|) \ \mathbf{P}-a.s.,\ n \to \infty. \tag{2.4}$$

Now the assumptions of Theorem 2.1.6 are fulfilled, which implies that the sequence $\{\varphi(|X_n^\omega|)\}$ is uniformly Leb-integrable ($\mathbf{P}$-a.s.). Now the equalities

$$\lim_{r\to\infty} \sup_n \int_{\varphi(|x|)>r} \varphi(|x|)dP_n^\omega = \lim_{r\to\infty} \sup_n \int_{\varphi(|X_n^\omega|)>r} \varphi(|X_n^\omega|)dLeb = 0 \ \mathbf{P}-a.s.$$

ensure that $\varphi(|x|)$ is uniformly integrable w.r.t. measures P_n ($\mathbf{P}$-a.s.).
The theorem is proved. □

2.4.2. Ergodic processes in practice

We will now present some examples of ergodic processes as models for various real life situations.

Example 2.4.10. Simplest example of ergodic process is a sequence of i.i.d. random variables $Y_1, Y_2, \ldots$. The stationarity and ergodicity of this process are well known.

Example 2.4.11 (Markov chain)**.** Remember that a sequence of random variables $\{X_n\}$ is called Markov chain, if

$$\mathbf{P}(Y_{n+1} \in B|\sigma(Y_1, \ldots, Y_n)) = \mathbf{P}(Y_{n+1} \in B|Y_n)$$

holds for all $B \in \mathcal{F}$. The Markov chain is determined by the initial distribution μ ($Y_0 \sim \mu$) and the transition density $p(x, A) = \sum_{y\in A} P(Y_{n+1} = y|Y_n = x)$. If the distribution μ is stationary, i.e. $\mu(A) = \int p(x, A)\mu(dx)$, then the sequence $Y_0, Y_1, \ldots$ is stationary as well. Assume that $\mu(x) > 0$ for all possible states x and the state space is countable. It can be proved that if the chain is irreducible (there is positive probability to move from any state to any other state with finite number of steps) then the σ-algebra of invariant elements is trivial, i.e. the Markov chain is ergodic. For details of this proof, see e.g. Durret (1991, pp. 293–294).

Stationary and ergodic processes have significant role in information theory and source coding theory. The objective of source coding theory is to achieve optimal performance of communication systems which must code an information source for transmission over a digital communication or storage channel for transmission to a user. The user must decode the information into a form that is a good approximation of the original. The optimal code has to give a most truthful picture with respect to the constraints (like rate or resolution) fixed by the channel. The information sources are mathematically modelled as discrete time random processes, and since the long behaviour of these sequences is of great interest, the ergodic theory can be applied.

In practice, there are several processes that behave similarly to stationary processes, although they do not strictly satisfy the definition. For example, processes can have nonstationarities due to transients that die out in time, and the distributions of samples will still converge to a stationary distribution as the sample times increase. This discussion results in defining the asymptotical mean stationarity.

Definition 2.4.12. A dynamical system $(\Omega, \mathcal{F}, \mathbf{P}, f)$ is called *asymptotically mean stationary* if there exists a stationary measure $\overline{\mathbf{P}}$ for which

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{P}(f^{-i}A) = \overline{\mathbf{P}}(A)$$

holds for each $A \in \mathcal{F}$.

Obviously every stationary system is also asymptotically mean stationary, with the original measure $\mathbf{P}$ being its asymptotical mean $\overline{\mathbf{P}}$. A known fact is that almost all information sources and processes produced by coding structures encountered in the real world are well modelled by asymptotically mean stationary processes. Often it is even reasonable to assume the strict stationarity of the process, since for any asymptotically mean stationary source there is an equivalent stationary source having same sample averages in long term. A complete survey of source coding theory is provided in Gray (1990).

Kokkuvõte

Tõenäosusjaotuste lähendamine hulkadega

Tõenäosusjaotuste lähendamine hulkadega on oluline ülesanne nii klassikalises statistikas ja tõenäosusteoorias kui ka mitmetes praktilistes valdkondades. Teatavasti on reaalteljel antud juhusliku suuruse X keskväärtus EX ja mediaan MeX parimad ühepunktilised lähendid vastavalt kaofunktsioonide $E(X-a)^2$ ja $E|X-a|$ mõttes. Käesolevas töös vaadeldakse selle klassikalise ülesande kaugeleulatuvat üldistust, seda nii lähendhulkade valiku, kaofunktsiooni kuju kui ka põhiruumi enda osas.

Olgu antud juhuslik element X jaotusega P separaablil meetrilisel ruumil (S,d) ja olgu $\mathcal{A}$ ruumi S teatud alamhulkade hulk. Lähendhulka $A \in \mathcal{A}$ nimetame optimaalseks (jaotuse P mõttes), kui ta minimiseerib järgmise kaofunktsiooni:

$$W(A,P) = \int_S \varphi(d(x,A))P(dx) \equiv E\varphi(d(X,A)),$$

kus $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ on mittekahanev hälbefunktsioon ja $d(x,A) = \inf\{d(x,a) : a \in \mathbb{R}\}$.
Võimalike lähendhulkade valik on väga lai. Tuntumad näited on k-punktilised hulgad, sirged, tasandid, aga ka erinevad kõverad ja pinnad. Kõige sügavamad tulemused on kirjanduses seni saadud juhul, kui klass $\mathcal{A}$ koosneb k-punktilistest hulkadest. Selliste lähendhulkadega on tegemist näiteks kvantimise (analoogsignaali muutmine digitaalseks) korral. Suhteliselt hästi on lahendatud optimaalsete k-hulkade olemasolu, nende koondumise jt. küsimused (vt. Pollard, 1981; Cuesta ja Matrán, 1988, 1989; Lember, 1999). Lähendhulkade koondumise probleem tekib siis, kui lähendhulgad A_n on optimaalsed mõõtude jada $\{P_n\}$ suhtes, kus $P_n \Rightarrow P$ (nõrgalt). Mainitud töödes on vaatluse all üksnes empiirilised mõõdud P_n ning väitekirja

üks ülesanne ongi käsitleda palju üldisemat juhtu, mis haaraks ka ergoodiliste protsesside poolt indutseeritud mõõte.

Meid huvitavad põhiküsimused on:

1) millal optimaalne lähendhulk üldse leidub?
2) kas kaofunktsioonide infiimumväärtuste jada $\{W(P_n)\}$ koondub?
3) kas (ε-)optimaalsete lähendhulkade jada $\{A_n\}$ koondub?

(Hulka A nimetame ε-optimaalseks jaotuse P jaoks, kui $W(A,P) \le W(P)+\varepsilon$.)

Antud probleeme on üsna laialdaselt uuritud ka varem (Pollard, 1981; Abaya ja Wise, 1984; Cuesta ja Matrán, 1988, 1989; Pärna, 1986, 1988; Averous ja Meste, 1997; Lember, 1999, 2002). Meie eesmärk on üldistada seniseid tulemusi antud valdkonnas kahes põhisuunas:

- valides võimalikult laia lähendhulkade klassi $\mathcal{A}$;
- tehes võimalikult vähe kitsendusi mõõtudele $\{P_n\}$ ja P, et kaasata ka praktikas olulisi mitte-empiirilisi mõõte.

Samal ajal on püütud teha võimalikult vähe kitsendusi ka hälbefunktsioonile φ ja põhiruumile S.

Töös kehtivad püsieeldused on toodud peatükis 2:

A1) φ on pidev, mittekahanev, Δ_2-omadusega, $\varphi(0)=0$ ja $\varphi(\infty)>0$;

A2) leidub $x_0 \in S$, mille korral $\int \varphi(d(x,x_0))P(dx) < \infty$;

A3) mõõtude jada $\{P_n\}$ koondub nõrgalt mõõduks P, st. $P_n \Rightarrow P$;

A4) leidub $x_0 \in S$, mille korral funktsioon $\varphi(d(x,x_0))$ on ühtlaselt integreeruv mõõtude jada $\{P_n\}$ suhtes.

Autoril õnnestus tõestada, et kahte viimast tingimust rahuldavad ka ergoodiliste protsesside poolt genereeritud mõõtude jadad (Teoreem 2.4.9).

Peatükis 3 on saadud mitu üldist tulemust suvalise lähendhulkade klassi ja separaabli meetrilise ruumi S korral. Esmalt on näidatud, et kõik ε-optimaalsed lähendhulgad lõikuvad piisavalt suure keraga (Teoreem 3.1.9). Seejärel on tõestatud (Teoreem 3.1.10), et üle selliste lähendhulkade klassi leiab aset ühtlane koondumine

$$\lim_n \sup_{\substack{A\in\mathcal{A} \\ A\cap\bar{B}(x_0,R)\neq\emptyset}} |W(A,P_n)-W(A,P)| = 0.$$

Vastus töö ühele põhiküsimusele – kaofunktsioonide optimaalsete väärtuste koondumine – on antud Teoreemis 3.1.11. Peatükis 3 tõestatud tulemused üldistavad mitmeid varasemaid samalaadseid tulemusi (vt. Pollard, 1981; Abaya ja Wise, 1984; Pärna, 1988; Lember, 1999).
Edasi on lähemalt uuritud lähendamist k hulga ühendiga, kus iga komponent pärineb klassist $\mathcal{A}$. Teoreemis 3.2.4 on näidatud, et piisavalt väikese $\varepsilon > 0$ korral leidub selline kera $\bar{B}(x_0, M)$, mis kaofunktsiooni $W(\cdot, P)$ suvalise ε-minimiseeriva jada $\{A_n(k)\}$ korral lõikub mittetühjalt selle *iga* komponenthulgaga. Neid üldisi tulemusi on kasutatud ülejäänud kahe põhiküsimuse – optimaalsete lähendhulkade olemasolu ja koondumise – uurimisel kahe lähendhulkade klassi korral. Enamus peatüki 3 tulemusi on avaldatud artiklis Käärik ja Pärna (2004).

Peatükis 4 on vaadeldud K-tõkestatud lähendhulkade klassi $\mathcal{A}^K$, s.t. sellist klassi, milles hulkade diameeter ei ületa konstanti K. Meid huvitab etteantud jaotuse lähendamine k hulgaga klassist $\mathcal{A}^K$. Sellised hulki võib käsitleda kui elemente meetrilises ruumis $(\overline{2}^S, h)$, mis on ruumi S kõikide kinniste mittetühjade alamhulkade ruum Hausdorffi kaugusega h. Kuna enamus ruumi S omadusi kanduvad üle ka ruumile $\overline{2}^S$, on ka uuritavad koondumistulemused ruumis $(\overline{2}^S, h)$ tugevalt seotud ruumi (S, d) omadustega. Teoreemis 4.4.3 on tõestatud optimaalsete lähendhulkade olemasolu. Teoreemis 4.4.4 on tõestatud P_n-ε-optimaalsete lähendhulkade jada koondumine P-ε-optimaalsete lähendhulkade klassi. Toodud tulemustest järeldub otse ka optimaalsete lähendhulkade koondumine (4.4.5). Saadud tulemused on edasiarenduseks artikli Käärik (2000) tulemustele, kus vaadeldi jaotuste lähendamist sfääridega. Peatüki lõpus on üldistusena uuritud ka jaotuste lähendamist k hulgaga erinevatest K-tõkestatud klassidest (Teoreem 4.5.2). Kõik need tulemused on palju üldisema iseloomuga võrreldes varasemate töödega selles vallas (vrdl. Pollard, 1981; Abaya ja Wise, 1984; Averous ja Meste, 1997; Pärna, Lember ja Viiart, 1999).

Peatükis 5 on käsitletud jaotuste lähendamist parameetriliste hulkadega. Iga lähendhulk on kirjeldatav kui k parameetrilise komponenthulga ühend ning iga selline ühend on määratud ühe (koond)parameetri Θ väärtusega parameetrite ruumist T^k. Kõigepealt on fikseeritud kolm olulist tingimust, mida 'hea' parametrisatsioon peab rahuldama:

- kujutis parameetrite ruumist T^k klassi $\mathcal{A}$ on lokaalselt ühtlaselt pidev, st kui parameetri väärtuste vaheline kaugus on väike, siis vastavate

lähendhulkade vaheline (Hausdorffi) kaugus on samuti väike iga kera piires ruumis S;

- parameetri väärtuste hulk ruumis T^k on tõkestatud, kui ruumis S leidub kera, mis lõikub kõigi vastavate komponenthulkadega;
- kui parameetri väärtuste hulk ruumis T^k on tõkestatud, siis ruumis S leidub kera, mis lõikub kõigi vastavate lähendhulkadega.

On eeldatud, et S on separaabel normeeritud ruum ja parameetrite ruum T on lõplikumõõtmeline normeeritud ruum. Tehniliste probleemide vältimiseks on esmalt vaadeldud juhtu, kus kõik komponenthulgad kuuluvad samasse hulkade klassi. Tõestatud põhitulemused on optimaalsete parameetri väärtuste leidumine (Teoreem 5.4.3) ja P_n-ε-optimaalsete parameetri väärtuste jada koondumine P-ε-optimaalsete väärtuste klassi (Teoreem 5.4.4). Nendest tulemustest järeldub ka optimaalsete parameetri väärtuste koondumine (Teoreem 5.4.5). Peatüki lõpus on näidatud, et tulemused jäävad kehtima ka situatsioonis, kus lähendhulga komponendid on pärit erinevatest klassidest (Teoreem 5.5.1). Enamus peatükis 5 toodud tulemustest on avaldatud artiklites Käärik ja Pärna (2003, 2004).

Bibliography

[1] Abaya, E.F., Wise, G.L. (1984). Convergence of vector quantizers with applications to optimal quantization. *SIAM Journal of Applied Mathematics*, **44**, 183–189.

[2] Averous, J., Meste, M. (1997). Median balls: an extension of the interquantile intervals to multivariate distributions. *Journal of Multivariate Analysis*, **63**, 222–241.

[3] Billingsley, P. (1968). Convergence of Probability Measures. Wiley, New York.

[4] Billingsley, P. (1995). Probability and Measure. Wiley, New York.

[5] Bock, H. H. (1996). Probabilistic models in cluster analysis. *Computational Statistics and Data Analysis*, **23**, 5–28.

[6] Chernov, N., Lesort, C. (2004). Least squares fitting of circles and lines. *Journal of Mathematical Imaging and Vision.* (to appear)

[7] Chernov, N., Ososkov, G. (1984). Effective algorithms for circle fitting. *Computer Physics Communications*, **33**, 329–333.

[8] Cuesta, J.A., Matrán, C. (1986). Strong Laws of Large Numbers in abstract spaces via Skorohod's Representation Theorem. *Sankhya*, **48**, 98–103.

[9] Cuesta, J.A., Matrán, C. (1988). The Strong Law of Large Numbers for k-means and best possible nets of Banach valued random variables. *Probability Theory and Related Fields*, **78**, 523–534.

[10] Cuesta, J.A., Matrán, C. (1989). Uniform consistency of r-means. *Statistics & Probability Letters*, **6**, 65–71.

[11] Cuesta-Albertos, J.A., Gordaliza, A., Matrán, C. (1998). On the geometric behaviour of multidimensional location measures. *Journal of Statistical Planning and Inference*, **67**, 191–208.

[12] Durrett, R. (1991). Probability: Theory and Examples. Duxbury Press, Belmont, California.

[13] Flury, B. (1990). Principal points. *Biometrika*, **77**, 33–41.

[14] Flury, B. (1993). Estimation of principal points. *Journal of the Royal Statistical Society: Series C*, **42**, 139–151.

[15] Gaenssler, P., Stute, W. (1979). Empirical processes: a survey of results for independent and identically distributed random variables. *The Annals of Probability*, **7**, 193–243.

[16] Graf, S., Luschgy, H. (2000). Foundations of Quantization for Probability Distributions. Springer-Verlag, Heidelberg.

[17] Gray, R. M. (1990). Source Coding Theory. Kluwer Academic Publishers, Boston-Dordrecht-London.

[18] Gray, R. M. (1991). Probability, Random Processes and Ergodic Properties. Springer-Verlag, New York.

[19] Gray, R. M., Neuhoff, D. L. (1998). Quantization. *IEEE Transactions on Information Theory* **44**, 2325–2383.

[20] Hartigan, J.A. (1978). Asymptotic distributions for clustering criteria. *The Annals of Statistics*, **9**, 117–131.

[21] Hastie, T., Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, **84**, 502–516.

[22] Käärik, M. (2000). Approximation of distributions by sphere. *Multivariate Statistics. New Trends in Probability and Statistics. Vol. 5*, 61–66. VSP/TEV, Vilnius-Utrecht-Tokyo.

[23] Käärik, M., Pärna, K. (2003). Approximation of distributions by parametric sets. *Acta Applicandae Mathematicae,* **78**, 175–183.

[24] Käärik, M., Pärna, K. (2004). Fitting parametric sets to probability distributions. *Acta et Commentationes Universitatis Tartuensis de Mathematica. Vol 8*, 101–112.

[25] Kégl, B., Krzyzak, A., Linder, T., Zeger, K. (2000). Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 281–297.

[26] Kendall, M. G., Stuart, A. (1966). The Advanced Theory of Statistics, Vol. 3: Design and Analysis, and Time Series. Griffin, London.

[27] Kuratowski, K. (1966). Topology. Vol. I. Mir, Moscow. (in Russian)

[28] Kuratowski, K. (1969). Topology. Vol. II. Mir, Moscow. (in Russian)

[29] Lember, J. (1999). Consistency of empirical k-centres. Doctoral Dissertation. Tartu University Press, Tartu.

[30] Lember, J. (2002). Consistency of k-centres via metric projection. *Limit Theorems in Probability and Statistics II*, Budapest, 335–350.

[31] Lember, J. (2003). On minimizing sequences for k-centres. *Journal of Approximation Theory*, **120**, 20–35.

[32] Lember, J., Pärna, K. (1999). Strong consistency of k-centres in reflexive spaces. *Probability Theory and Mathematical Statistics*. VSP/TEV, Vilnius-Utrecht-Tokyo, 441–452

[33] Lloyd, S.P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, **28**, 129 – 136.

[34] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 281–297.

[35] Michael, E. (1951). Topologies on spaces of subsets. *Transactions of the American Mathematical Society*, **71**, 152– 182.

[36] Oja, E., Oja, P. (1991). Funktsionaalanalüüs. Tartu University Press, Tartu. (in Estonian).

[37] Pärna, K. (1986). Strong consistency of k-means clustering criterion in separable metric spaces. *Acta et Commentationes Universitatis Tartuensis*, **733**, 86 – 96.

[38] Pärna, K. (1988). On the stability of k-means clustering criterion in separable metric spaces. *Acta et Commentationes Universitatis Tartuensis*, **798**, 19 – 36.

[39] Pärna, K. (1990). On the existence and weak convergence of k-centres in Banach spaces. *Acta et Commentationes Universitatis Tartuensis*, **893**, 17 – 28.

[40] Pärna, K., Lember, J. and Viiart, A. (1999). Approximating of distributions by sets. *Classification in the Information Age*, 215 – 224, Springer-Verlag, Heidelberg.

[41] Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics*, **9**, 135–140.

[42] Pollard, D. (1984). Convergence of Stochastic Processes. Springer-Verlag, New York-Berlin-Heidelberg-Tokyo.

[43] Ranga Rao, R. (1962). Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics*, **33**, 659 – 680.

[44] Shiryayev, A. N. (1980). Probability. Nauka, Moscow. (in Russian)

[45] Späth, H. (1996). Least-squares fitting by circles. *Computing*, **57**, 179–185.

[46] Späth, H. (1997a). Least-squares fitting of ellipses and hyperbolas. *Computational Statistics*, **12**, 329–341.

[47] Späth, H. (1997b). Orthogonal distance fitting by circles and ellipses with given area. *Computational Statistics*, **12**, 343–354.

[48] Späth, H. (1997c). Orthogonal least squares fitting by conic sections. *Recent Advances in Total Least Squares techniques and Errors-in-Variables Modeling*, SIAM, 259–264.

[49] Sverdrup-Thygeson, H. (1981). Strong Law of Large Numbers for measures of central tendency and dispersion of random variables in compact metric spaces. *The Annals of Statistics*, **9**, 141 – 145.

[50] Tarpey, T., Flury, B. (1996). Self-consistency: a fundamental concept in statistics. *Statistical Science*, **11**, 229–243.

[51] Tibshirani, R. (1992). Principal curves revisited. *Statistics and Computation*, **2**, 183–190.

[52] Vapnik, V. (1998). Statistical Learning Theory. Wiley, New York.

[53] Varadarajan, V.S. (1958). On the convergence of probability distributions. *Sankhya*, **19**, 23–26.

[54] Verbeek, J. J., Vlassis, N., Kröse, B. (2002). A k-segments algorithm for finding principal curves. *Pattern Recognition Letters*, **23**, No. 8, 1009–1017.

[55] Verbeek, J. J., Vlassis, N., Kröse, B. (2001). A soft k-segments algorithm for principal curves. *Artificial Neural Networks - ICANN 2001. Lecture Notes in Computer Science*, **2130**, 450–456. Springer-Verlag, Heidelberg.

[56] Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, **10**, 299–326.

Curriculum Vitae

Name: Meelis Käärik
Citizenship: Estonian Republic
Born: November 3, 1976, Tartu, Estonia
Marital Status: married, 1 child
Address: J. Liivi 2-503, Insitute of Mathematical Statistics, University of Tartu
Contacts: e-mail: Meelis.Kaarik@ut.ee

Education

1994–1998: Faculty of Mathematics, University of Tartu, BSc in mathematical statistics (*cum laude*)
1998–2000: Faculty of Mathematics, University of Tartu, MSc in mathematical statistics
2000–2004: Faculty of Mathematics and Computer Science, University of Tartu, PhD studies in mathematical statistics

Professional employment

1995–2004: Computer Administrator, Institute of Mathematical Statistics, University of Tartu
2000–2003: Project Manager, Resta Ltd.
2004– : Researcher, Institute of Mathematical Statistics, University of Tartu

Curriculum Vitae

Nimi: Meelis Käärik
Kodakondsus: Eesti Vabariik
Sünniaeg ja -koht: 3. november 1976, Tartu, Eesti
Perekonnaseis: abielus, 1 laps
Aadress: J. Liivi 2-503, Tartu Ülikooli matemaatilise statistika instituut
Kontaktandmed: e-mail: Meelis.Kaarik@ut.ee

Hariduskäik
1994–1998: Tartu Ülikooli matemaatikateaduskond, bakalaureusekraad matemaatilise statistika erialal (*cum laude*)
1998–2000: Tartu Ülikooli matemaatikateaduskond, magistrikraad matemaatilise statistika erialal
2000–2004: Tartu Ülikooli matemaatika-informaatikateaduskond, doktoriõpingud matemaatilise statistika erialal

Erialane teenistuskäik
1995–2004: arvutivõrgu administraator, Tartu Ülikooli matemaatilise statistika instituut
2000–2003: projektijuht, AS Resta
2004– : teadur, Tartu Ülikooli matemaatilise statistika instituut